March 12, 2018

CS 361: Probability & Statistics

Inference

Binomial likelihood: Example

Suppose we have a coin with an unknown probability of heads. We flip the coin 10 times and observe 2 heads. What can we say about the probability of heads?

Binomial likelihood

In N independent coin flips, we observe k heads. What is the maximum likelihood estimate for θ ?

We want to calculate

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

What is the likelihood function in this setup?

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Binomial likelihood

We need to take a derivative and set equal to 0

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \binom{N}{k} \left[k\theta^{k-1}(1-\theta)^{N-k} - (N-k)\theta^k (1-\theta)^{N-k-1} \right]$$

Set $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ and solve for theta

$$k\theta^{k-1}(1-\theta)^{N-k} = (N-k)\theta^k (1-\theta)^{N-k-1}$$

Binomial likelihood

$$k\theta^{k-1}(1-\theta)^{N-k} = (N-k)\theta^{k}(1-\theta)^{N-k-1}$$

giving

$$k(1-\theta) = (N-k)\theta$$

or

$$k - k\theta = N\theta - k\theta$$

Thus our max likelihood estimate is given by

$$\hat{\theta} = \frac{k}{N}$$

Geometric likelihood

Suppose we flip a coin, stopping after we see a head. We do this and observe that it took us N flips. What is the maximum likelihood estimate of θ?

Once again, we want to compute

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

This time, we have

$$\mathcal{L}(\theta) = (1-\theta)^{N-1}\theta$$

Geometric likelihood

We want to choose a theta to maximize L

$$\mathcal{L}(\theta) = (1-\theta)^{N-1}\theta$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = (1-\theta)^{N-1} - (N-1)(1-\theta)^{N-2}\theta$$

Taking $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ we get

$$(N-1)(1-\theta)^{N-2}\theta = (1-\theta)^{N-1}$$

$$\theta N - \theta = 1 - \theta$$

Giving an MLE estimate of $\hat{\theta} = \frac{1}{N}$

Log likelihood

Since the logarithm is a monotonic function, maximizing the log of the likelihood will give us the same theta as if we maximize the likelihood

$$\arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \log \mathcal{L}(\theta)$$

Logarithms allow us to break products up into sums which will make some computations easier

$$\log P(\mathcal{D}|\theta) = \log \prod_{i \in \text{dataset}} P(d_i|\theta) = \sum_{i \in \text{dataset}} \log P(d_i|\theta)$$

Poisson likelihood: example

Suppose we work in the maternity ward of a large hospital and we begin to write down how many babies are born each hour and get the following dataset

Hour	# of babies
1	4
2	2
3	0
4	1
5	3
6	0
7	1
8	0
9	1
10	3
11	0
12	1
13	2
14	1
15	1

If we suppose the number of babies born each hour is governed by a Poisson distribution, what is the maximum likelihood estimate for its intensity, given the data?

Poisson likelihood

Suppose we observe some event of interest over N intervals each of the same fixed length and that the number of events that we observe in interval i is given by n_i. We model this situation with a Poisson random variable and we want to know the max likelihood estimate for λ

Getting the derivative of our likelihood might be tricky since we have

$$\mathcal{L}(\theta) = \prod_{i \in \text{intervals}} P(\{n_i \text{ events}\}|\theta) = \prod_{i \in \text{intervals}} \frac{\theta^{n_i} e^{-\theta}}{n_i!}$$

But we can easily maximize

$$\log \mathcal{L}(\theta) = \sum_i (n_i \log \theta - \theta - \log n_i!)$$

Poisson likelihood

Let's differentiate the following with respect to theta

$$\log \mathcal{L}(\theta) = \sum_i (n_i \log \theta - \theta - \log n_i!)$$

to get

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \sum_i \left(\frac{n_i}{\theta} - 1 \right)$$

Setting $\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = 0$ we get $\frac{1}{\theta} \left(\sum_i n_i \right) - N = 0$

Solving for theta we get our MLE $\hat{\theta} = \frac{\sum_i n_i}{N}$

Example

Suppose we work in the maternity ward of a large hospital and we begin to write down how many babies are born each hour and get the following dataset

Hour	# of babies
1	4
2	2
3	0
4	1
5	3
6	0
7	1
8	0
9	1
10	3
11	0
12	1
13	2
14	1
15	1

If we suppose the number of babies born each hour is governed by a Poisson distribution, what is the maximum likelihood estimate for its intensity, given the data?

On the last slide we had

$$\hat{\theta} = \frac{\sum_i n_i}{N}$$

So our MLE for λ is 20/15

Normal likelihood

Assume we observe *N* data items x_1, x_2, … , x_N thought to conform to a normal distribution. What is the max likelihood estimate for the mean of this normal distribution given the data?

We have

$$\mathcal{L}(\theta) = P(x_1, x_2, \ldots, x_N | \theta, \sigma)$$

Or

$$\mathcal{L}(\theta) = P(x_1|\theta, \sigma)P(x_2|\theta, \sigma)\ldots P(x_N|\theta, \sigma)$$

Which if we use the distribution of the normal, gives us

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

Normal likelihood

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

would be a pain to work with, so we look at the log-likelihood given by

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \theta)^2}{2\sigma^2}\right) + \text{term not depending on } \theta$$

Differentiating with respect to theta, we get

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{2(x_i - \theta)}{2\sigma^2}$$

Normal likelihood

Setting the derivative of the log likelihood equal to 0 we get

$$0 = \sum_{i=1}^{N} \frac{2(x_i - \theta)}{2\sigma^2}$$

We get

$$N\theta = \sum_{i=1}^{N} x_i$$

Simplifying

$$0 = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \theta\right)$$

Giving an MLE of

$$\hat{\theta} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Normal likelihood

Suppose we have N data items as before but want a maximum likelihood estimate for the standard deviation of the normal distribution our data are coming from

This time we have

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\theta} \exp\left(-\frac{(x_i - \mu)^2}{2\theta^2}\right)$$

So our log likelihood is given by

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\theta^2}\right) - N \log \theta + \text{term not depending on } \theta$$

Normal likelihood

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\theta^2} \right) - N \log \theta + \text{term not depending on } \theta$$

differentiating with respect to theta we get

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{-2}{\theta^3} \sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2} - \frac{N}{\theta}$$

Simplifying and setting equal to 0

$$0 = \sum_{i=1}^{N} (x_i - \mu)^2 - N\theta^2$$

For an MLE of

$$\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Maximum likelihood: drawbacks

A couple of things might trip up max likelihood estimation:

1) Finding the maximum of some functions can be quite hard

2) If we don't have a large amount of data, we might incorrectly estimate certain model parameters

For example, the MLE for p in a binomial distribution is k/N if we have observed k heads in N coin flips. If we have observed zero heads in 2 flips a coin, is it always safe to assume $p=0$?

Bayesian inference

Bayesian inference

An alternative method for doing parameter estimation — figuring out a good theta, given the data — that has a different set of strengths and weaknesses than maximum likelihood estimation is called Bayesian inference

With MLE, we tried to find a theta that maximized the likelihood function

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

With Bayesian inference, we maximize a different function of theta, by treating theta as a random variable

$$P(\theta|\mathcal{D})$$

The value of theta that maximizes this function is called the **maximum a posteriori estimate** or **MAP estimate**

The prior

From Bayes' rule, we know that we can express our function of interest as

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

The right hand side contains the likelihood, which we've been working with. Also in the numerator is the so-called **prior probability** of θ

Bayesian inference is useful because it allows us to incorporate prior beliefs we have about the value of θ

The prior

From Bayes' rule, we know that we can express our function of interest as

Likelihood Prior

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

In principle we can use any distribution we want for the prior, if we wanted to stubbornly insist that the true value of θ must not be between 0.25 and 0.75 no matter what the data tells us, for instance, we can just have a prior that has a probability of 0 for those values of θ and our MAP estimate will never be in that range

The prior

From Bayes' rule, we know that we can express our function of interest as

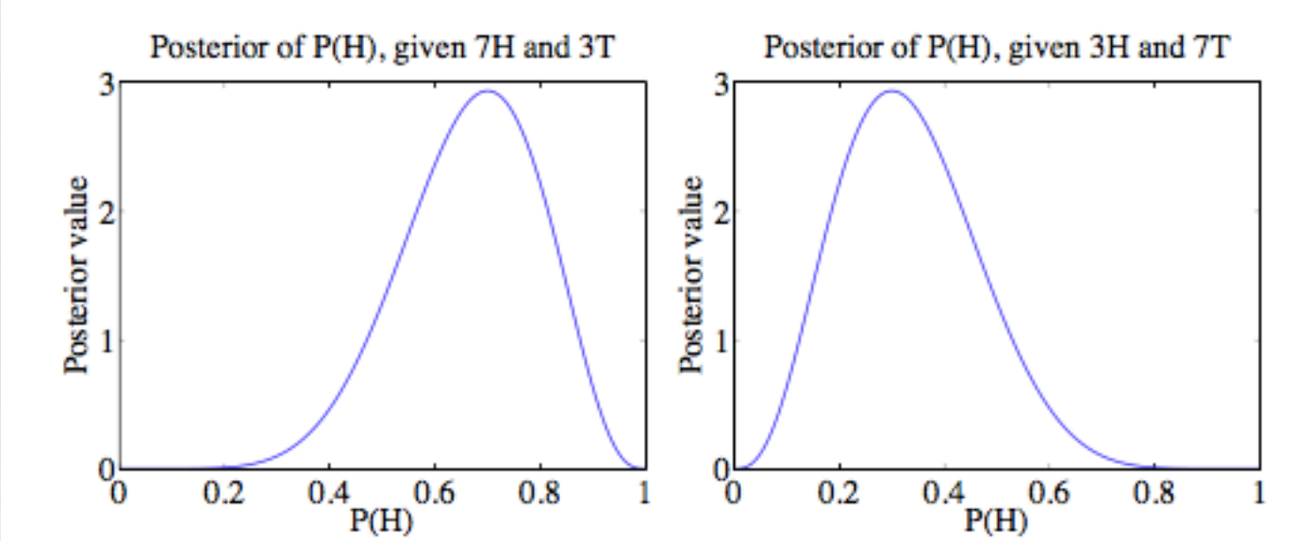
$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(\mathcal{D})}$$

If we had a uniform prior, on the other hand, we are saying we have no particular beliefs about the true value of θ

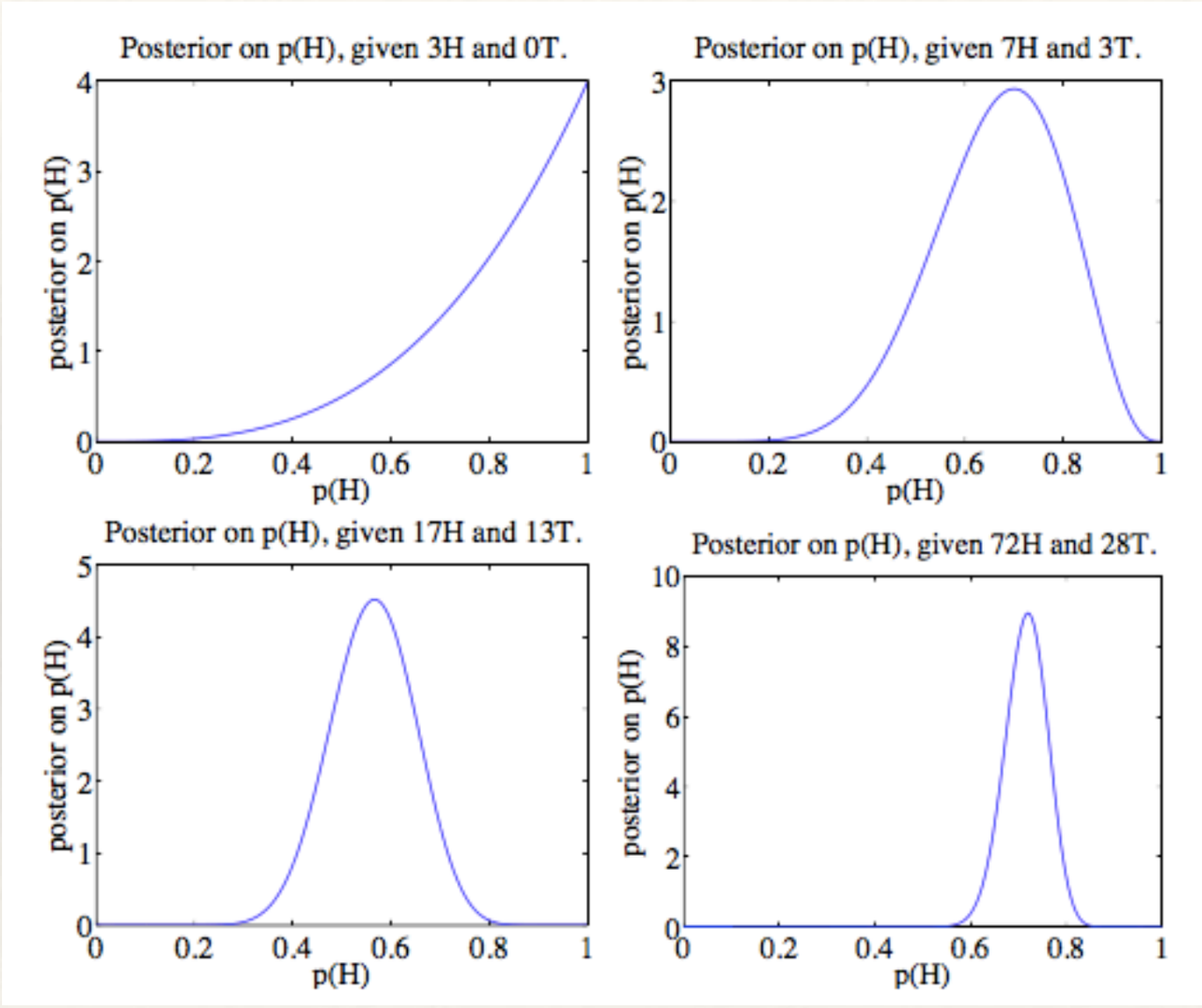
In that case, choosing a theta to maximize the left hand side (the MAP estimate) is the same as choosing a theta to maximize the likelihood (the MLE estimate) since only the likelihood on the RHS depends on theta

Example

Suppose we have a coin with probability θ of heads coming up. We make no assumptions about the prior probability of θ (i.e. assume a uniform prior). We flip the coin 10 times and see 7 heads and 3 tails. Plot a function proportional to p(θ | 7 heads and 3 tails) and for 3 heads, 7 tails.



Example


Posterior on p(H), given 3H and 0T.
posterior on p(H)
p(H)
Posterior on p(H), given 7H and 3T.
posterior on p(H)
p(H)
Posterior on p(H), given 17H and 13T.
posterior on p(H)
p(H)
Posterior on p(H), given 72H and 28T.
posterior on p(H)
p(H)

Which prior?

So we are in this interesting situation where we are considering the probability of a probability in some sense. The probability that a coin is fair or that it comes up heads 90% of the time.

In order to have a good prior for the theta in a coin flipping Binomial model we need a function that is a probability density over the range [0,1]

When we are doing a MAP estimate, we are trying to maximize the product

$$P(\mathcal{D}|\theta)P(\theta)$$

We want this product to be well-behaved enough that we can optimize it to get our MAP estimate, and multiplying two different probability distributions together might produce a really ugly result

Which prior?

Likelihood Prior

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Posterior

A particular kind of good behavior we might insist upon is the following:

1) For a given problem setup, the likelihood function is largely out of our control. E.g. we suppose that our data is from a normal distribution, the likelihood function is going to be normal
2) So the prior is our only degree of freedom
3) We choose a prior that is expressive enough that we can encode arbitrary beliefs about the prior probability of theta — the unknown parameters in our model
4) But choose a prior such that when it is multiplied by the likelihood function, we get a posterior that is of the same random variable type as the prior

A prior satisfying 4) above is called a **conjugate prior** of the likelihood function

Which prior, binomial

The binomial family of distributions is conjugate to the **beta** family of distributions

A beta random variable is a continuous random variable defined on $0 \leq x \leq 1$ with parameters alpha > 0 and beta > 0 whose density has the following form

$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}(1-x)^{\beta-1}$$

The constant is in terms of a special function called the **gamma function** which is a generalization of the factorial function to positive real values rather than just non-negative integers. Details can be found in the first chapter of the book

$$p(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

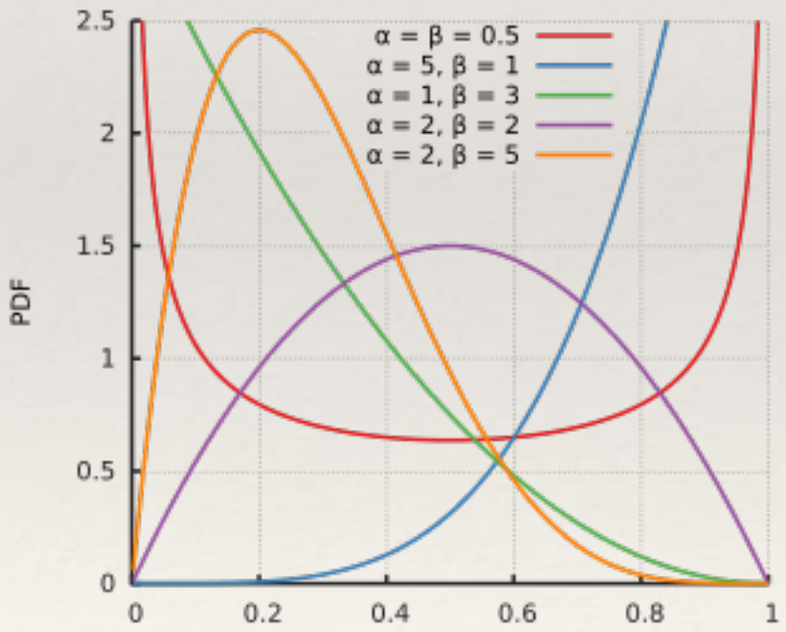
Beta distribution

Useful Facts: 6.6 *Beta distribution*

For a Beta distribution with parameters α, β

1. The mean is $\frac{\alpha}{\alpha+\beta}$.
2. The variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

$$p(x;\alpha,\beta) = (\text{constant})x^{\alpha-1}(1-x)^{\beta-1}$$



The beta distribution is very expressive

Having alpha=beta=1 would give a uniform prior

Binomial likelihood, beta prior

So if we want to do Bayesian inference against a binomial problem setup. Our likelihood be a binomial distribution. Let's see what happens if we pair that likelihood with a beta prior

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

If our data in this case is that we observed h heads in N flips, we have

$$p(\theta|\mathcal{D}) \propto \binom{N}{h}\theta^h(1-\theta)^{N-h}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Likelihood

Prior

Just focusing on theta, we get

$$p(\theta|\mathcal{D}) \propto \theta^{\alpha+h-1}(1-\theta)^{\beta+N-h-1}$$